International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-07, Issue-10, JAN 2022

Sports Analytics – Soccer/Football Player Market Value Analysis and Potential Prediction

[1]Zaid Asif Basri, [2]Anirudh Parasuraman Iyer, [3]Karthik Sivaraman Iyer, [4]Dr. Deepti Reddy
[1,2,3]UG Student, [4]Professor, SIES Graduate School of Technology – Navi Mumbai, INDIA,
[1]basrizaid3108@gmail.com, [2]anirudh.iyer1@gmail.com, [3]karthikiyer365@gmail.com,
[4]deepti.reddy@siesgst.ac.in

Abstract: In Soccer/Football, scouting is one of the most important processes for player recruitment. Football clubs invest millions of dollars in signing players who would be the best fit for their team every year. This process requires scouts to have analytical and observational skills. Scouts have to analyze the player on their actions, physical attributes, age, position of play, and other attributes before making a judgement regarding the player. Teams have a formation, a style of play and the player requirements for any year and position may change irregularly. A scout's judgment of any player based solely on observation from a couple of matches cannot be considered as accurate. This process is flawed as the scout is expected to watch a few games and make estimates of how well the player would fit in their side. There is a lack of using the player's performance statistics, career statistics and some physical attributes in making these decisions. An imperfect analysis, leading to unreliable scouting would cause a loss of millions to the club and its owners. Our proposal is to solve the problem of imperfect scouting by using performance statistics from the years past and present, incorporating data science techniques and ML models to make educated estimates. We take into account criterion including position of play, defensive abilities, offensive abilities, player's play style, suitable formations etc. The addition of scientific analysis of data would help in better understanding the players strength, benefits and weaknesses, thereby facilitating scouts in making better decisions.

Keywords — Joint Offensive Impact, Joint Defensive Impact, Regression Analysis, Random Forest Algorithms, Event stream data, , Naïve Bayes classifier, Soccer Player Action Description Language, Support Vector Machines (SVMs), Valuing Action by Estimating Probabilities

I. INTRODUCTION

Football is a widely played sport, with players of different caliber across the globe. Any team in the world aims to have the best possible players in their team but is only able to shed a limited amount of monetary funds on each player. The process of analyzing players from around the world and finding the most suitable player for a particular position in a team is called scouting. This is a crucial part in the world of football since getting the right player is of utmost importance. Only specific kinds of players would contribute toward the success of the team in the various competitions they participate in. The player recruitment process is overseen by individuals who physically attend a few matches of a potential new player. They are known as scouts. Scouts observe the player, collect data, and evaluate the talent of the player.

Scouting is such an important aspect of the footballing world, that an official Football Scout Association was formed in 2013 with the aim of providing professional scouting courses. This association helps to learn how to analyze players, not only on number of goals scored but also on the contribution of player actions towards the various stages of the team's overall play. Attempts to automate the process of understanding player's gameplay, team coherence as well as many other attributes have been made. Wyscout, ScoutMe, Sciports and many more such professionals have tried to and have established a method to automate the process of scouting. The systems developed have been fairly accurate in establishing a basic sense of the crucial players. Our primary goal is to realize a process and system that would incorporate an understanding of these crucial aspects, thereby suggesting players that suit the tactical setup of any given team.

Our system would incorporate feature engineering, to display an understanding of the different set of player traits that shine through in the different positions in football. These traits come from an understanding of the style of play, tactical formation of the team, the current line-up, playing style of the team, vacant positions etc. This data is analysed by our algorithm, which in turn returns the list of

 DOI : 10.35291/2454-9150.2022.0021

the top players in accordance with the tactical attributed of the payer. The list also is calculated in accordance to the spending limit of each team on a basis of a price tag calculated for them. This data is fed to an algorithm which in turn returns the list of top players, according to physical, tactical and fiscal criteria of that position. The list then given as input to our ML model which will analyze the player's past data and return a list of the top N players. Thus, our system will aim to present results with some ground truth and help scouts to estimate player worth, or automate the process of procuring players, with better accuracy in estimating the impact of any player in a team

II. Literature Review

Abbreviations: Soccer Player Action Description Language (SPAD), Valuing Actions by Evaluating Probabilities.

The most complex operation in soccer data analysis is to process game data and transform it into useful information. This information is crucial as it represents the first step in player scouting, and furthermore leading to signing new players for team development. Going through the existing research in this domain, we come across the structure of analysis adopted and implemented.

The initial problem of understanding player performance is referred to and tackled by utilizing two frameworks. SPADL, which is a language for representing soccer data, and VAEP, which is a framework that records the actions of any player, help in mining the real-time video feed to generate a database of players and a value for each attribute of their game [2]. The implementation of these help to minimize on apparent human error, increase efficiency as well as accuracy of player analysis. Using wyscout data, the author explains the usuage of SPADL and VAEP [10].

Furthermore, a different approach to data mining is crowd sourcing which is discussed in multiple papers and undermined [4]. In their paper, Oliver Müller, Alexander Simons, Markus Weinmann, discuss the inefficacy yet imply the need for crowd sourcing methods [11]. Data analytics proves to be a sound alternative compared to traditional crowdsourcing methods. But the paper outlines that mass opinion is not to be considered null and void. The data driven approach is most optimum when considering attributes of gameplay, but is not accurately able to understand market value of players without taking into consideration popularity of players. The paper gives us insight into how crowdsourcing and data driven analysis work complimentarily and not alternatively.

The study conducted to establish the relationship between performance and transfer price for football players points out that profit is secondary when it comes to player transfers. The European Big 5 leagues are success-maximizers as opposed to profit-maximizers [1]. This research shows that teams focus on transferring better players, which further boosts their market value as clubs, compete to sign said player. A checklist of potential transfers would comprise of factors like number of months remaining on an existing contract, their gameplay and cross positioned compatibility, all of which are easily analyzed through crowdsourcing and data driven methods. Cross positioned compatibility is discussed as a means to understand the diversity in positions that a player can perform at [8]. But the list is topped by the potential of a player and their ability to merge with the clubs existing formation.

As a means to understand the potential of a player and their compatibility with existing team, we sought after papers discussing team style sports, not exclusively soccer. Referring to such sports, we came across the need to analyze team game play. Gameplay chemistry can be understood to be better if two players have played together earlier in any form of the sport [7]. This is better understood by analyzing game feed from any players history and understanding their most interacted plays. The more any two players interact, the better their chemistry, whether it be in training, while representing their nations, or in any different form of the sport. Considering a factor of potential, we make use of SPADL to see their growth within and across seasons. Considering this, they can be classified over a scale. To understand the valuing of a player we referred to player valuation techniques using a data mining approach. A decision tree approach, with factors as mentioned earlier, is considered while valuing players [6].

Overall, we understand the various data mining approaches that shines light on the need to not exclude crowdsourcing. The various factors like popularity and potential that are generally missed out are highlighted. Our review helped us to understand that valuations for any physical entity with a well-defined set of parameters is a challenge in itself and depends on numerous other things. And doing the same for an individual with varying parameters, based on the source of data, is even more challenging. This paper is an attempt to find the most optimum way of fetching data of football players and applying the right model on it so as to extract meaningful information, thus reducing the gap between the estimated prices and the final price [9].

III. Problem Statement

PROPOSED SYSTEM

The proposed system is to design a model that would look at a player's past performances and based on an analysis done, the player would be rated on a scale of 1-100. The model would also look to predict this player's value for the coming years. Thus, according to the needs of a team/club, they can invest in the right player and make sure that their

 DOI : 10.35291/2454-9150.2022.0021

investments bring them great returns. The objective is to create a model that looks to calculate player's attributes and assign a market value for teams to invest based on their requirements and desires.

PROPOSED SYSTEM FOR EXISTING DRAWBACKS

It is no secret that player performance analysis helps a team/club make the necessary changes to their way of playing the game. Using statistical analysis for the entire team's playing style as well as for keeping track of the strengths and weaknesses of each player are all new ways for teams to catch up with the game to incorporate a win. That being said, it is just as important for a team/club to have the mentality of having a vision or to think about the future of the team.

With the increase in competitiveness in the sports field, a high demand for accurate & descriptive data has become a necessary aspect to analyze. In football, transfer markets play one, among many roles for a team in order to succeed. Hence, hiring the right set of personnel and signing the right players are equally essential. Football clubs spend a huge amount of money every year to buy professional football players, during the transfer window. Predicting how much a certain player values in the transfer market is one of the difficult tasks for managers of the club. For example, a certain player may be at the peak of his career. At that point, without having any foresight, pundits would call him the best player in the world. However, using the statistics to determine whether the player would be able to achieve similar results in the coming years and whether he would be able to cope up with the dynamics of the game with a different club. If yes, then what would be the right price to sign him? Using this as a thought process, we look to design a model that would help the team management decide which player would be fit for the club and what would be the right price to sign him.

To achieve the above, we propose a Three-fold framework

1. Phase 1 – Evaluation of players based on current form
2. Phase 2 – Prediction of a player's form for the coming seasons
3. Phase 3 – Identification of players based on team/club requirement

The system would require certain inputs from the team/club management. Based on those inputs, our system would identify the type of players demanded by the club and give back the output in the form of a list. The primary objective of our project is to create a model that understands a player's attributes and assign a certain value along with a proper price-tag (market value) of that player.

- **COMPONENTS FOR PROPOSED SYSTEM**

SPADL and VAEP framework:

Our dataset is based on the collection of many sequential actions performed by individual players. The data that we are using is present in JSON format which is not easy to visualize and process. Thus, we use SPADL to convert the complex json format data to readable format. VAEP is used to grade each action based on the probability value calculated by SPADL. These actions that are graded, are all in-game actions performed. The advantages of SPADL and VAEP are:

1.1. Using SPADL, we can store every action of an individual player and thus it becomes easy to understand and process the data for further analysis.

1.2. Through VAEP, based on the action performed, we can assign a positive score to each player when the action performed is successful and a negative score when the action has failed.

OBJECTIVES

The primary objective is to create a model that encapsulates and understands the dynamics of the game and understands a player's attributes and assigns a certain value and a proper price-tag to that player. The model aims to help clubs and teams analyze players based on their age and playing skills and based on that use, our model identifies the right price for that player. This helps in making sure that the negotiation is fair and also benefits the cub as it gives insights about the players future.

- This would help the team or club target the right player for them.
- This would help the team or club to avoid incurring any losses due to overpaying a certain individual as the market value (in terms of currency) for the player will be set.

DESIGN & DATASETS

Datasets:

We have used data from various sources to cover all aspects of the player, which includes physical attributes, actions, tactics etc.

The datasets that we have used in our system are as follows:

FIFA

This is a collection of 6 csv files of which five include player attributes and statistics and one is all about the teams and the leagues in which the teams/clubs play. The five FIFA player csv files include 18,279 rows which are the number of players and 104 columns which give us the number of attributes that differentiate a player from another. The motivation behind choosing this dataset was

 DOI : 10.35291/2454-9150.2022.0021

that the historical data from 2015 to 2020 enables us to study the progression of players and the respective growth in their key attributes. The attributes can further be streamlined into 6 major categories based on each player and their specialty (position they play)

IV. METHODOLOGY

The project's main objective is to find out suitable players based on the team/club's requirement. Before coming to a conclusion or a list of players, at the backstage, we carry out another procedure to provide better suggestions to the team. The model would evaluate each player's past performance and rate him on a scale of 1-100. After this, the model would also calculate the future rating or potential rating of that player. After all these calculations, the model would also suggest a price tag for his current rating and would also depict as to whether the player's price would correspondingly increase, decrease, or stay constant. For easier understanding, we can divide the same as a Three-fold framework. This would help in micromanaging each process and ensure efficiency and smooth execution.

I. Phase 1 – Evaluation of the players based on past and current form

The primary objective of this stage is to evaluate each player and assign a rank (based on the rating). In order to achieve this, there are two tasks that need to be done.

1. Using SPADL and VAEP to evaluation of actions and rating attributes on a scale of 1-100 for each player.

1.1. SPADL:

- SPADL stands for Soccer Player Action Description Language. As the name suggests, it talks about describing a player's actions. That means, giving the machine an idea about the player's movements.
- Now to do this we use two forms of actions. Either Event stream data or Optical or Visual tracking data.
- This procedure is carried out by making note of actions performed by players. Based on the success-rate (determined on the basis of how many times the player is able to execute that action successfully), the probability is determined for each action.
- This value illustrates how consistent a player can perform a particular action successfully in terms of probability.

1.2. VAEP:

- VAEP stands for Valuing Actions by Estimating Probabilities.
- VAEP is a framework in python itself and is a part of SPADL.
- It makes use of the probability values we obtain from SPADL and uses it to calculate a rating.

1.3. List of Actions Rated:

Pace	Physical
GK Diving	GK Handling
GK Reflexes	GK Speed
Positioning	Crossing
Finishing	Header
Short pass	Long pass
Vision	Volleys
Freekick accuracy	Ball control
Shot Power	Shot Accuracy
Interceptions	Take on
Sliding Tackle	Mentality

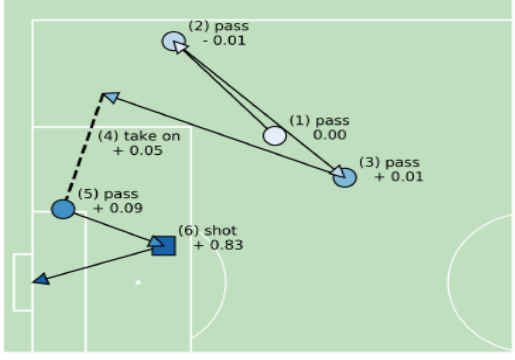
1.4. Determining Market Value

Based on the rating calculated, each rating has a specific Market value associated with it. Using this value, we determine the current price-tag of all the players.

2. Categorization and clustering of players based on attributes, position, age and few other parameters.

- All the players are then categorized into 4 major categories viz. Forwards, Mid-Fielders, Defenders, Goalkeepers.
- The players are all rated based on the ratings for each action and are then arranged in a particular manner.
- In order to get this task done, we make use of Random Forest Algorithms. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many Decision trees.

	TIME	PLAYER	ACTION	P_{scores}	VALUE
1	92m4s	S. Busquets	pass	0.03	0.00
2	92m6s	L. Messi	pass	0.02	- 0.01
3	92m8s	S. Busquets	pass	0.03	+ 0.01
4	92m11s	L. Messi	take on	0.08	+ 0.05
5	92m12s	L. Messi	pass	0.17	+ 0.09
6	92m14s	A. Vidal	shot	1.00	+ 0.83



Example of evaluation of player actions (VAEP) [10]

 DOI : 10.35291/2454-9150.2022.0021

The image above depicts a scenario from a football/soccer match in the year 2018 between FC Barcelona and Real Madrid. As we can see, the table consists of multiple attributes and observations. The first column shows the time at which the action took place during the match. Followed by that, we can see the name of the player performing the action, the action (whether it was a pass, a shot, a take on etc.). We can then look at the probability - this quantifies the success of the action leading to a goal. The Value column indicates how much the said action is contributing towards either scoring or conceding a goal. Using this value, VAEP is able to calculate how consistently a player can successfully complete an action in order to score maximum goals and concede minimum. The formula to calculate the same is as follows:

$$V(a_i,x) = \Delta P_{scores}(a_i,x) + (-\Delta P_{concedes}(a_i,x))[10]$$

II. Phase 2 – Prediction of a player's form (potential) for the future

The objective of this phase is to calculate the potential of the player. Along with the potential, to provide details of the future that include attributes as well as to provide the details of market value will be done in this phase.

1. Calculation of Potential Rating:

a. The calculation of the potential is done by taking into account thorough data of the past that ranges from datasets of the football season 2015-16 to 2020-21.

b. On equipping our model with these datasets, we make use of Regression Analysis to obtain the most accurate output.

c. Support Vector Machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection.

d. The model also makes use of Naïve Bayes classifiers which works very fast and can easily predict the class of a test dataset. It is used to solve multi-class prediction problems as it's quite useful with them. Naive Bayes classifier performs better than other models with less training data if the assumption of independence of features holds.

2. Calculation of Future Market Value of Player:

As we have calculated the present price-tag of players using their rating, we take into account two factors that would help in calculating the potential market value of the corresponding players.

First, the rise in rating. This would show how much or how consistent the player has been. The rise in potential would obviously mean a hike in the market value of that player. However, a descent or plateau in the performance of the player would result in a subsequent decline in market value.

Second, the time in which the potential has increase. Taking age into account it is but natural that an older player even though might have strong data in the past to back him up with a good potential, he may not be having the same level of consistency as his age increases. Similarly, a younger player may not have enough data backing him, but he has many years to bring his potential up.

Thus, based on the two factors above, we make use of Linear Regression Analysis to figure out the potential rating and corresponding potential market value of the players.

Name (Pos)	*Pace*	*Shooting*	*Passing*	*Dribbling*	*Physical*	*OVR*
Messi (RW)	85	92	91	95	65	93
Ronaldo (ST)	89	93	81	89	77	91

(a)

Name (Pos)	*Pace*	*Defending*	*Passing*	*Dribbling*	*Physical*
Van Dijk (CB)	78	91	71	72	84
Ramos (CB)	70	88	76	74	84

(b)

Evaluated attributes of from EA FIFA *[12]*
(a) Lionel Messi and Cristiano Ronaldo
(b) Virgil van Dijk and Sergio Ramos

EA Sports - FIFA 22 is a game that virtually encapsulates the entire football/soccer fraternity. The game is able to provide realistic data of players. In the game, to keep it as realistic as possible, EA makes use of huge chunks of datasets of players and teams. As a result, we can get a fair idea of the ranking of players based on their performance.

SPADL is a method that makes use of the VAEP values obtained previously and converts it into the overall attribute rating of each player based on each action rated. It is calculated using the following formula:

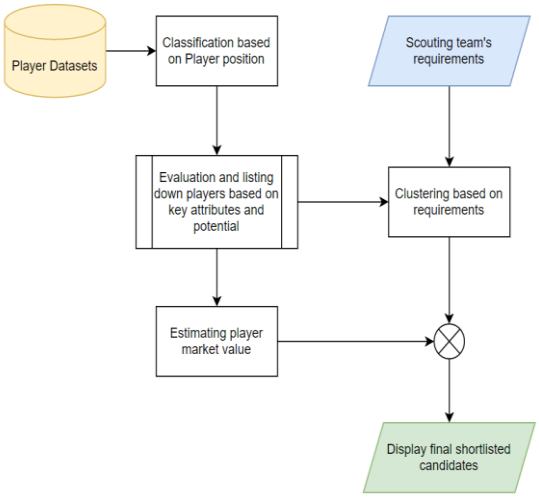
$$rating(p) = 90/m \; S \; V(a_i) \ldots\ldots where \; a_i \; \epsilon \; A^T_p \; [10]$$

III. Phase 3 – Identification of players based on team/club requirement

The model will make use of a user-friendly web-page that would help the team management easily get their desired results/outputs and thus gain insights. In order to suggest some players, the model makes use of CatBoost and XG Boost without any explicit pre-processing to convert categories into numbers. CatBoost converts categorical values into numbers using various statistics on combinations of categorical features as well as

 DOI : 10.35291/2454-9150.2022.0021

combinations of categorical and numerical features. These methods assist in the recommendation of players.



Flowchart depicting the flow of processes carried out by proposed model

V. Plans For Next Segment

In the three months spent on this project, we have gone through various existing systems as well as understood the different types of models developed to do football player analysis and team development. The existing systems demonstrate a thorough analysis of the players attributes and team requirements. Our system will aim to conduct a similar analysis of the dataset established by the SPADL and VAEP models.
Our next few steps are as follows: - To clean and filter the data that has been generated using VAEP and SPADL models

- To develop a model to rank players based on their attributes and action values.
- To establish a monetary value for the players in accordance to their ranking, possibility of transfer as well as potential.
- To analysis team requirements and generate a list of possible transfers.
- To develop a website to represent the possible player transfers and data representation

VI. Conclusion

Our system takes into account a multitude of statistics of every player's all-round game, and combines the quantitative and qualitative aspects to generate a list of players that are of a very specific type and fit in a specific style of playing. We then use this list to simulate the impact that the player will have on the current team, thereby giving the manager an educated estimate of what he/she can expect from the newly recruited player, which can potentially save the club a huge amount of money, compared to following traditional scouting methods which operate without any detailed extrapolation of the incoming player's impact based on current performance.

VII. Applications & Future Scope

The project has a wide range of applications starting right from local clubs at school levels, to the top leagues in world football. Scouting is an inherent process in every team. The system is built so as to accept the requirement from the coach or scout looking for players. The system will convert the data from the various platforms into one common language known as SPADL, then it calculates the action values of the players based on this database created. These action values, VAEP values, are matched with the player requirement given by the team to the system. We consider the actions of the players and not the goal scoring ability. This helps us in objectively analyzing players at all levels of the sport. The project can be taken forward by broadening the scope on introducing real-time data as well as incorporating concepts of Joint Offensive Impact (JOI) and Joint Defensive Impact (JDI). Furthermore, we develop a monetary understanding of the players considering factors like their age and potential to shine in upcoming seasons. Since the system gives a thorough understanding of the players as well as considers factors from the current team, it is widely applicable for all teams. We perform a thorough analysis of the previous teams, the current state of the team before predicting a possible transfer. This is required for teams at all leagues and since this is a real time analysis, the system can be used for a better approach toward any teams scouting.

The project can be used to identify young and hidden talent for football clubs. The system can be used by all international, national and local level teams. Teams such as Chelsea, Manchester United, Manchester City, Liverpool can use the system in analyzing player across other clubs and signing new players. National teams of India, Argentina, Germany etc. can use the system in analyzing gameplay video of new upcoming talents from within the nation. This will help the national teams in upgrading the status and uplifting the level of gameplay. Hence, our system helps the cubs too save a huge amount of money which they lose if they hire imperfect players. Therefore, our application plays an important role in modern day player scouting and helps clubs save money, as well as developing the team thoroughly.

References

[1] M Jean-Sébastien Jacques Emile Marie Ghislain Geurts, 751 "Football players' transfer price determination based on performance in the Big 5 European leagues

[2] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan and I. Matthews, "Identifying Team Style in Soccer Using Formations Learned from Spatiotemporal

 DOI : 10.35291/2454-9150.2022.0021

Tracking Data" 2014 IEEE International Conference on Data Mining Workshop, 2014, pp. 9-14, doi: 10.1109/ICDMW.2014.167.

[3] Coutinho J.C., Moreira J.M., de Sá C.R. (2020) "UnFOOT: Unsupervised Football Analytics Tool." In: Brefeld U., Fromont E., Hotho A., Knobbe A., Maathuis M., Robardet C. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Lecture Notes in Computer Science, vol 11908. Springer, Cham.

[4] "Real-Time Data Acquisition and Performance Analysis in Sports" Hristo Novatchkov, Sebastian Bichler, Martin Tampier, Philipp Kornfeind and Arnold Baca, Department of Biomechanics, Kinesiology and Computer Science, Faculty of Sport Science, University of Vienna, Auf der Schmelz 6A, 1150 Vienna, Austria.

[5] "Comprehensive Data Analysis and Prediction on IPL using Machine Learning" Algorithms Amala Kaviya V.S.1, Amol Suraj Mishra2 and Valarmathi B.3 1Member of Technical Staff - Grade 2, VMware India Pvt. Ltd., Bangalore (Karnataka), India. 2Member of Technical Staff - Grade 2, NetApp, Bangalore (Karnataka), India. Associate Professor, Department of Software and Systems Engineering, School of Information Technology and Engineering, Vellore Institute of Technology, Vellore (Tamilnadu), India.

[6] Kansal, P., Kumar, P., Arya, H., & Methaila, A. (2014). "Player valuation in Indian premier league auction using data mining technique." 2014 International Conference on Contemporary Computing and Informatics (IC3I). doi:10.1109/ic3i.2014.7019707.

[7] Lotte Bransen, Jan Van Haaren, "Player Chemistry: Striving for a Perfectly Balanced Soccer Team".

[8] Nazim Razali, Aida Mustapha, Faiz Ahmad Yatim, Ruhaya Ab Aziz, "Predicting Player Position for Talent Identification in Association Football".

[9] Dibyanshu Patnaik, Harsh Praharaj, Kartikeya Prakash, Prof. Krishna Samdani, NMIMS University, "A study of Prediction models for football player valuations by quantifying statistical and economic attributes for the global transfer market", Proceeding of International Conference on Systems Computation Automation and Networking 2019.

[10] Tom Decroos, Lotte Bransen, Jan Van Haaren, Jesse Davis, "Actions Speak Louder than Goals: Valuing Player Actions in Soccer".

[11] Oliver Müller, Alexander Simons, Markus Weinmann, European Journal of Operational Research 263 (2017) 611-624, "Beyond crowd judgements: Data-driven estimation of market value in association football"

[12] https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset

List of Figures:

Sr.No.	Figure Description
1.	Example of evaluation of player actions (VAEP) [11]
2.	Evaluated attributes of Cristiano Ronaldo(left) and Sergio Ramos(right) from EA FIFA [14]
3	Flowchart depicting the flow of processes carried out by proposed model

 DOI : 10.35291/2454-9150.2022.0021